Differentiable Probabilistic Models

William Watson
nextbillyonair@gmail.com

Abstract

d

Contents

4.1 Mixture Model

4.2 Gumbel Mixture Model

4.3 Infinite Mixture Model

5 Transforms

Transforms are invertible functions that can be applied to a random variable to change the distribution.

5.1 Transform

5.2 Inverse Transform

5.3 Affine

- **Parameters**
  - Location $\mu \in \mathbb{R}^n$
  - Scale $\sigma > 0$
- **Forward**

$$f(x) = \mu + \sigma \cdot x \tag{1}$$

- **Inverse**

$$f^{-1}(y) = \frac{y - \mu}{\sigma} \tag{2}$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = \log |\sigma| \tag{3}$$

5.4 Exp

- **Parameters**
  - None
- **Forward**

$$f(x) = e^x \tag{4}$$

- **Inverse**

$$f^{-1}(y) = \log y \tag{5}$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \tag{6}$$

5.5 Expm1

- **Parameters**
  - None
- **Forward**

$$f(x) = e^x - 1 \tag{7}$$

- **Inverse**

$$f^{-1}(y) = \log(1 + y) \tag{8}$$

- **Log Absolute Determinant Jacobian**

$$\log |\det \mathbf{J}|(x, y) = x \tag{9}$$

5.6 Gumbel

- **Parameters**
  - Location $\mu \in \mathbb{R}^n$
  - Scale $\sigma > 0$
- **Forward**

$$f(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) \tag{10}$$

- **Inverse**

$$f^{-1}(y) = \mu - \sigma \cdot \log\left(-\log\left(y\right)\right) \tag{11}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -\log\left(\frac{\sigma}{-\log(y) \cdot y}\right) \tag{12}$$

5.7 Log

- **Parameters**
  - None
- **Forward**

$$f(x) = \log x \tag{13}$$

- **Inverse**

$$f^{-1}(y) = \exp y \tag{14}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -y \tag{15}$$

5.8 Logit

- **Parameters**
  - None
- **Forward**

$$f(x) = \log\left(\frac{x}{1-x}\right) \tag{16}$$

- **Inverse**

$$f^{-1}(y) = \frac{1}{1+e^{-y}} \tag{17}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = \log\left(1+e^{-y}\right) + \log\left(1+e^{y}\right) \tag{18}$$

5.9 Power

- **Parameters**
  - Power p
- **Forward**

$$f(x) = \begin{cases} e^x & p = 0 \\ (1 + x \cdot p)^{1/p} & \text{otherwise} \end{cases} \tag{19}$$

- **Inverse**

$$f^{-1}(y) = \begin{cases} \log y & p = 0 \\ y^{p-1}/p & \text{otherwise} \end{cases} \tag{20}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = \begin{cases} x & p = 0 \\ \left(\frac{1}{p} - 1\right) \cdot \log\left(x \cdot p + 1\right) & \text{otherwise} \end{cases} \tag{21}$$

5.10 Reciprocal

- **Parameters**
  - None
- **Forward**

$$f(x) = 1/x \tag{22}$$

- **Inverse**

$$f^{-1}(y) = 1/y \tag{23}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -2 \cdot \log|x| \tag{24}$$

5.11 Sigmoid

- **Parameters**
  - None
- **Forward**

$$f(x) = \frac{1}{1 + e^{-x}} \tag{25}$$

- **Inverse**

$$f^{-1}(y) = \log\left(\frac{y}{1 - y}\right) \tag{26}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -\log\left(1 + e^{-x}\right) - \log\left(1 + e^{x}\right) \tag{27}$$

5.12 SinhArcsinh

5.13 Softplus

5.14 Softsign

5.15 Square

5.16 Tanh

6 Criterion and Divergences

The criterion and divergences listed here can be used to quantify the "distance" between two distributions. Hence, in conjunction with torch optimizers, one can minimize said difference to learn the paramters of a distribution. For sake of notation clarity, p is the true distribution and q is the learned distribution. Hence we "fit" q to match p. In addition, we provide the Monte Carlo approximation.

6.1 Cross-Entropy

$$\begin{aligned} H(p, q) &= -\int p(x) \log q(x) dx \\ &= -\frac{1}{n} \sum_{x \sim p} \log q(x) \end{aligned} \tag{28}$$

6.2 Perplexity

$$\begin{aligned} H(p, q) &= \exp\left(-\int p(x) \log q(x) dx\right) \\ &= \exp\left(-\frac{1}{n} \sum_{x \sim p} \log q(x)\right) \end{aligned} \tag{29}$$

6.3 Forward KL Divergence

$$
\begin{aligned}
H(p,q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\
&= \frac{1}{n} \sum_{x \sim p} \log \frac{p(x)}{q(x)}
\end{aligned}
\tag{30}
$$

6.4 Reverse KL Divergence

$$
\begin{aligned}
H(p,q) &= \int q(x) \log \frac{q(x)}{p(x)} dx \\
&= \frac{1}{n} \sum_{x \sim q} \log \frac{q(x)}{p(x)}
\end{aligned}
\tag{31}
$$

6.5 Jensen-Shannon Divergence

6.6 Earth Mover's Distance

7 ELBO

8 Adversarial Loss

Adversarial Losses are criterion functions that allow for sample-sample based training between models p and q. More formally, it hides a Discriminator model that attempts to discriminate between the real data from p and fake data generated from q.